A two-class discrete-time queueing model with two dedicated servers and global FCFS service discipline

Herwig Bruneel Willem Mélange Bart Steyaert Dieter Claeys
Joris Walraevens

Department of Telecommunications and Information Processing
Ghent University - UGent
E-mail: {hb,wmelange,bs,dc,jw}@telin.UGent.be

Abstract

This paper considers a simple discrete-time queueing model with two types (classes) of customers (types 1 and 2) each having their own dedicated server (server A and B resp.) New customers enter the system according to a general independent arrival process, i.e., the total numbers of arrivals during consecutive time slots are i.i.d. random variables with arbitrary distribution. Service times are deterministically equal to 1 slot each. The system uses a "global FCFS" service discipline, i.e., all arriving customers are accommodated in one single FCFS queue, regardless of their types. As a consequence of the "global FCFS" rule, customers of one type may be blocked by customers of the other type, in that they may be unable to reach their dedicated server even at times when this server is idle, i.e., the system is basically non-workconserving. One major aim of the paper is to estimate the negative impact of this phenomenon on the queueing performance of the system, in terms of the achievable throughput, the system occupancy, the idle probability of each server and the delay. As it is clear that customers of different types hinder each other more as they tend to arrive in the system more clustered according to class, the degree of "class clustering" in the arrival process is explicitly modelled in the paper and its very direct impact on the performance measures is revealed. The motivation of our work are systems where this kind of blocking is encountered, such as input-queueing network switches or road splits.

Key words: queueing; non-workconserving; blocking; dedicated servers; global FCFS

1 Introduction

In general, queueing phenomena occur when some kind of customers, desiring to receive some kind of service, compete for the use of a service facility (containing one or multiple servers) able to deliver the required service. Most queueing models assume that a service facility delivers exactly one type of service and that all customers requiring this type of service are accommodated in one common queue. If more than one service is needed, multiple different service facilities are provided, i.e., one service facility for each type of service, and individual queues are formed in front of these service facilities. In all such models, customers are only hindered by other customers that require exactly the same kind of service, i.e., that compete for the same resources.

In some applications, it may not be physically feasible or desirable to provide separate queues for each type of service that customers may require, and it may be necessary or desirable to accommodate different types of customers (i.e., customers requiring different types of service) in the same queue. In such cases, customers of one type (i.e., requiring a given type of service) may also be hindered by customers of other types. For instance, if a road or a highway is split in two or more subroads leading to different destinations, cars on that road heading for destination A may be hindered or even blocked by cars heading for destination B, even when the subroad leading to destination A is free, simply because they have to queue in first-come-first-served (FCFS) order on the main road. This blocking also takes place in weaving sections on highways [16, 17] and in left-turn traffic models [25]. We refer to [22, 23] for a general overview and validation of the modelling of traffic flows with queueing models. Analogously, at a security checkpoint at for instance international airports or trainstations, people have to be bodysearched by someone of the same gender. As a result, when a group of friends of the same gender arrive, the people of the opposite gender behind them may have to wait until the whole group is checked, even when the other security person is available, when it is not allowed to overtake at the security checkpoint (for security reasons). Similarly, in switching nodes of telecommunication networks, information packets with a given destination A may have to wait for the transmission of packets destined to node B that arrived earlier, even when the link to destination A is free, if the arriving packets are accommodated in so-called input queues according to the source from which they originate (the well-known HOL-blocking effect, see [1, 10, 18, 19, 26, 27]). These situations are also related to models where queues are "pooled" (see e.g. [15, 21]) in the sense that customers (cars or packets) that require a different service or have a different destination share a common queue. Although these queues can be considered as pooled, the difference with the models in [15, 21] is that customers can be blocked by customers of the other type.

In order to gain insight into the impact of this kind of phenomenon on the performance of the involved systems, we study a simple conceptual discrete-time queueing model in this paper, which is simple enough to allow explicit solution but rich enough to capture the essential aspects of the problem at hand.

The structure of this paper is as follows. In section 2, the model under investigation is described in detail. Section 3 first presents a general analysis of the number of customers in the system: an expression is derived for the pgf of this number and a method is described

to determine the two remaining unknowns in that expression. Next, for the special case of geometric arrivals, explicit closed-form expressions are obtained not only for the pgf but also for the pmf and the mean value of the number of customers in the system. In section 4, we turn to the idle probability of each server. Section 5 is devoted to the study of the customer delay: by introducing complex contour integration, we succeed in deriving explicit expressions for the pgf of the delay in terms of the pgf of the number of customers in the system, applicable for any type of arrival distribution. As in section 3, the special case of geometric arrivals leads to considerable simplifications and allows for an explicit derivation of the pmf and the mean value of the customer delay. We discuss the results both conceptually and quantitatively in section 6. Finally, some conclusions are drawn in section 7.

2 Mathematical model

We consider a discrete-time queueing system with infinite waiting room, two servers, named A and B, and two types (classes) of customers, named 1 and 2. Each of the two servers is dedicated to a given class of customers, i.e., server A can only serve customers of type 1 and server B can only serve customers of type 2. Service times of all customers are deterministically equal to 1 slot each. Customers are served in their order of arrival, regardless of the class they belong to. We call this service discipline "global FCFS" in this paper.

The arrival process of new customers in the system is characterized in two steps.

First, we model the total (aggregated) arrival stream of new customers by means of a sequence of i.i.d. discrete random variables with common probability mass function (pmf) $e(n)$ and common probability generating function (pgf) $E(z)$ respectively. More specifically,

$$e(n) \triangleq \text{Prob}[n \text{ arrivals in one slot}] \quad , \quad n \geq 0 \ ,$$

$$E(z) \triangleq \sum_{n=0}^{\infty} e(n)\, z^n \ .$$

The total mean number of arrivals per slot, in the sequel referred to as the mean arrival rate, is given by

$$\lambda = E'(1) \ .$$

Next, we describe the occurrence of the two types (1 and 2) in the sequence of the consecutive arriving customers. We assume that both types of customers account for half of the total load of the system, i.e., both customer classes are equiprobable, but there may be some degree of "class clustering" in the arrival process, i.e., customers of any given type may (or may not) have a tendency to "arrive back-to-back". Mathematically, this means that the types of two consecutive customers may be non-independent. Specifically, we assume a first-order Markovian type of correlation between the types of two consecutive customers, which basically means that the probability that the next customer belongs to a given class depends on the class of the previous customer. We denote by α the probability that the next customer has *the same type as the previous one*, and by $1 - \alpha$ the probability that the next customer belongs to *the*

opposite type as the previous one. The parameter α can then be considered as a measure of the degree of class clustering in the arrival process, and will therefore be referred to as the "cluster parameter" in the sequel. It is easily seen that the size of a cluster of customers of the same type, i.e., the number of consecutive customers of any given type between two customers of the opposite type, is geometrically distributed with parameter α and mean value $1/(1-\alpha)$.

It can be seen that the two-server system described above is non-workconserving, for two different (orthogonal) reasons. First, the fact that the two servers A and B are dedicated to only one type of customers each, may result in situations where only one of the servers is active even though the system contains more than one customer (of the same type, in such a case). This implies that we cannot expect the system to perform as well as a regular two-server queue with two equivalent servers, i.e., servers able to serve *all* customers. In this paper, we consider this form of inefficiency as an intrinsic feature of our system, simply caused by the fact that the customers as well as the servers are non-identical. The second reason why the system is non-workconserving lies in the use of the global FCFS service discipline. This rule may result in situations where only one server is active although the system contains customers of *both* classes. Such situations occur whenever the two "eldest" customers in the system, i.e., the two customers at the front of the queue, are of the same type: only one of them can then be served (by its own dedicated server) and the other "blocks" the access to the second server for customers of the opposite type further in the queue. This second form of inefficiency is not an intrinsic feature of two-class systems with dedicated servers, but rather it is due to the accidental order in which customers of both types happen to arrive (and receive service) in the system (as described by the parameter α in our model). It is this second mechanism that we want to emphasize in the paper. For this reason, we have considered single-slot service times and equiprobable customer classes. These assumptions make the system completely symmetric, in the sense that the number of customers that can be served during a slot does not depend on the actual type of the customer in head-of-line position, but only on the identity or non-identity of the classes to which the two customers at the front of the queue belong. This symmetry ensures that the obtained formulas are elegant, so that they reveal the very direct and great influence of the degree of class clustering in the arrival stream (via parameter α) on the stability and the main performance measures of the system.

It is worth noting that this model can be conceived as a model with batch service [4], whereby the service threshold is set equal to 1 (i.e., as soon as 1 customer is present, a new batch service is initiated), and whereby the server capacity, i.e., the maximum number of customers that can be served simultaneously, is a random variable (with probability α equal to 1, and with probability $1-\alpha$ equal to 2, provided that at least 2 customers are awaiting service in the system). Batch-service queueing models with a variable service capacity are quite difficult to analyze, and as far as we know no relevant results exist in the literature from which the results in our paper can be derived.

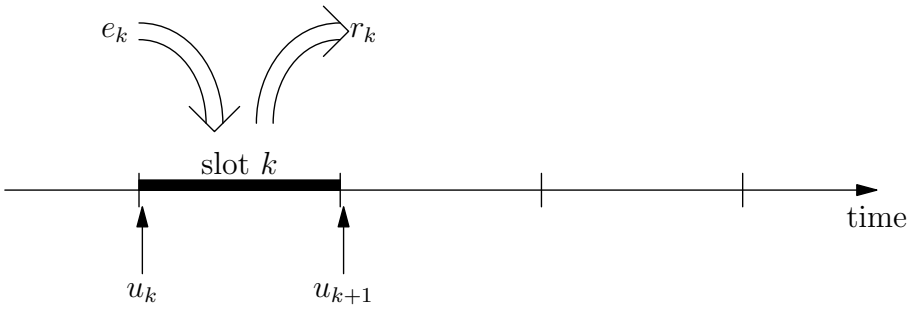

Figure 1: Time axis to illustrate the system equations

3 System occupancy

3.1 System equations

We start the analysis by defining a number of important random variables, illustrated in Fig. 1. Specifically, let u_k denote the total system occupancy, i.e., the total number of customers present in the system at the beginning of the k-th slot, and e_k the total number of arrivals in the system during this slot (with known pmf $e(n)$ and pgf $E(z)$). Furthermore, let r_k (initially) denote the number of customers served during the k-th slot, when $u_k > 1$. Then the following recursive system equations can be established:

$$u_{k+1} = e_k \ , \quad \text{if } u_k \leq 1 \ ,$$
$$u_{k+1} = u_k + e_k - r_k \ , \quad \text{if } u_k > 1 \ .$$

The two above cases can be summarized in one single system equation

$$u_{k+1} = e_k + (u_k - r_k)^+ \ , \tag{1}$$

by introducing the notation $(\ldots)^+$ to indicate the quantity $\max(0, \ldots)$.

In equation (1), the random variables $\{r_k\}$ can be treated as a sequence of strictly positive i.i.d. random variables (indicating the numbers of "available servers" during the consecutive slots) with common pmf

$$r(n) \triangleq \text{Prob}[r_k = n] \ , \quad 1 \leq n \leq 2 \ ,$$

and common pgf

$$R(z) \triangleq \sum_{n=1}^{2} r(n)\, z^n \ ,$$

whereby

$$r(1) = \alpha \ , \quad r(2) = 1 - \alpha \ ,$$

and

$$R(z) = \alpha z + (1 - \alpha) z^2 \ . \tag{2}$$

In fact, this observation is the key to the solution. It actually turns out that the number of customers that can be served in slot k (with $u_k > 1$) does not depend on the actual type of the customer in the head-of-line position, but only on the identity or non-identity of the classes to which the two "eldest" customers (at the front of the queue) belong, regardless of the numbers of customers served during previous slots. If both customers belong to the same class, which happens with probability α, irrespective of the type of the head-of-line customer, then only one customer can be served. If the two customers belong to opposite classes, then both will be served; this case occurs with probability $1 - \alpha$. This explains why $r(1) = \alpha$ and $r(2) = 1 - \alpha$. It is clear that equation (1) is also correct if $u_k \leq 1$, because, with the given definition of the r_k's, $(u_k - r_k)^+$ is equal to zero in such cases.

3.2 Analysis of the system occupancy

For all k, let $U_k(z)$ denote the pgf of u_k. Then, from equation (1) we can derive

$$U_{k+1}(z) = E(z) \cdot E\left[z^{(u_k - r_k)^+}\right] \quad , \tag{3}$$

with $E[\cdot]$ the expectation operator. The second factor in the right hand side of (3) can be expanded further by means of the law of total probability (using also the mutual independence of u_k and r_k):

$$E\left[z^{(u_k - r_k)^+}\right] = \alpha E\left[z^{(u_k - 1)^+}\right] + (1 - \alpha) E\left[z^{(u_k - 2)^+}\right] \quad . \tag{4}$$

Here, the two remaining expectations are to be taken with respect to one single random variable u_k. Using standard z-transform techniques in equation (4), and combining the result with (3), we then obtain

$$U_{k+1}(z) = E(z) \cdot \left(R(1/z) U_k(z) + \frac{z - 1}{z^2} \left[(z + 1 - \alpha) u_k(0) + (1 - \alpha) z u_k(1) \right] \right) \quad , \tag{5}$$

where, for all $i \geq 0$,

$$u_k(i) \triangleq \text{Prob}[u_k = i] \quad .$$

Now, let us assume that the queueing system at hand is stable, i.e., that the stability condition is fulfilled. It is not difficult to see that, with the system equations established above, the system is stable if and only if the mean number of arrivals per slot, given by $E'(1)$, is strictly less than the mean number of available servers per slot, given by $R'(1)$, i.e., if and only if

$$E'(1) < R'(1) \quad ,$$

or, expressed in the basic parameters of our system,

$$\lambda < 2 - \alpha \quad . \tag{6}$$

We now let the time parameter k go to infinity. Assuming the system reaches a steady state, then both functions $U_k(z)$ and $U_{k+1}(z)$ converge to a common limit function $U(z)$, which

denotes the pgf of the system occupancy at the beginning of an arbitrary slot in steady state. As a result, equation (5) translates into a linear equation for $U(z)$, with solution

$$U(z) = \frac{(z-1)E(z)\,[u(0)(z+1-\alpha) + u(1)(1-\alpha)z]}{z^2 - (1-\alpha+\alpha z)E(z)} \ , \tag{7}$$

where

$$u(i) \triangleq \lim_{k\to\infty} u_k(i) \ .$$

This expression contains only known quantities, except for the two unknown probabilities $u(0)$ and $u(1)$. These can be determined, in general, by invoking the well-known property that pgf's such as $U(z)$ are bounded inside the closed unit disk $\{z : |z| \leq 1\}$ of the complex z-plane, at least when the stability condition (6) of the queueing system is met (only in such a case our analysis was justified and $U(z)$ can be viewed as a legitimate pgf). Now, it can be shown by means of Rouché's theorem from complex analysis [8, 3] that the denominator of equation (7) has exactly two zeroes inside the closed unit disk of the complex z-plane, one of which is equal to 1, as soon as the stability condition (6) is fulfilled. It is clear that these two zeroes should also be zeroes of the numerator of equation (7), as $U(z)$ must remain bounded in those points. For the zero $z = 1$, this condition is fulfilled regardless of the values of the unknowns $u(0)$ and $u(1)$, since the numerator of (7) contains a factor $z - 1$. However, for the second zero, the requirement that the numerator should vanish yields a linear equation for the two unknowns. A second linear equation can be obtained by invoking the normalizing condition of the pgf $U(z)$, i.e., the condition $U(1) = 1$. In general, the two unknown probabilities $u(0)$ and $u(1)$ can be found as the solutions of the two established linear equations. Substitution of the obtained values in equation (7) then leads to a fully determined expression of the steady-state pgf $U(z)$ of the system occupancy.

From this result, various performance measures of practical importance can then be derived. For instance, the mean system occupancy can be found as $E[u] = U'(1)$. By applying (the discrete-time version of) Little's result [12, 3, 5], the mean delay (system time) of a customer can be obtained as $E[d] = U'(1)/\lambda$, and so on. In the next subsection, we treat a special case in which the computations can be further simplified and explicit closed-form expressions can be obtained for most quantities of interest.

3.3 Special case: geometric arrivals

Let us consider the special case whereby the number of arrivals per slot has a geometric distribution with mean value λ. Then, $e(n)$ and $E(z)$ are given by

$$e(n) = \frac{1}{1+\lambda}\left(\frac{\lambda}{1+\lambda}\right)^n \quad , \quad n \geq 0 \ ,$$

$$E(z) = \frac{1}{1+\lambda-\lambda z} \ ,$$

and (7) can be rewritten as

$$U(z) = \frac{u(0)(z+1-\alpha) + u(1)(1-\alpha)z}{-\lambda z^2 + z + (1-\alpha)} \ , \tag{8}$$

where we have cancelled out a common factor $z-1$ from the numerator and the denominator.

It is not difficult to see that, as soon as the stability condition (6) is satisfied, the (quadratic) denominator of (8) has two zeroes, one of which (z_1) is inside the unit disk, and one of which (z_0) is outside the unit disk. As explained above, the bounded nature of $U(z)$ inside the unit disk implies that z_1 should also be a zero of the numerator of equation (8), which happens to be a linear function of z. It then follows that $U(z)$ can be further simplified by cancelling out the common factor $z - z_1$ from the numerator and the denominator and using the normalizing condition $U(1) = 1$. As a result we obtain

$$U(z) = \frac{1-z_0}{z-z_0} \ , \tag{9}$$

where z_0 is given by

$$z_0 = \frac{1+\sqrt{1+4\lambda(1-\alpha)}}{2\lambda} \ . \tag{10}$$

The pgf $U(z)$ given in equation (9) can be easily inverted; the corresponding pmf of the steady-state system occupancy reads

$$u(i) = \left(1 - \frac{1}{z_0}\right)\left(\frac{1}{z_0}\right)^i \quad , \quad i \geq 0 \ , \tag{11}$$

i.e., the system occupancy has a geometric distribution with parameter $1/z_0$.

The tail distribution Prob$[u > i]$, i.e., the probability that more than i customers be present in the system — which can be used as a rough approximation for the loss probability in a finite-capacity system with room for exactly i customers, see [20, 9, 11] — can be expressed as

$$\text{Prob}[u > i] = \left(\frac{1}{z_0}\right)^{i+1} \quad , \quad i \geq 0 \ . \tag{12}$$

The mean system occupancy $E[u]$ at the beginning of an arbitrary slot can be easily derived as well:

$$E[u] = \frac{1}{z_0 - 1} = \frac{1-2\lambda-\sqrt{1+4\lambda(1-\alpha)}}{2(\lambda-2+\alpha)} \ . \tag{13}$$

Finally, the mean delay $E[d]$ of a customer (expressed in time slots) can be obtained from the discrete-time version of Little's result [3, 5]:

$$E[d] = \frac{E[u]}{\lambda} = \frac{1-2\lambda-\sqrt{1+4\lambda(1-\alpha)}}{2\lambda(\lambda-2+\alpha)} \ . \tag{14}$$

It is worth noting that the stability condition (6) is clearly reflected in the expressions (13) and (14), in that the denominators of both expressions tend to zero as the mean arrival rate λ approaches its limiting value $2-\alpha$, indicating the unbounded growth of (mean) buffer occupancy and delay as the system approaches the border of its stability region.

4 Idle Probability

In this section, we deduce the idle probability for each server. The first server can be idle in three cases:

- If the system is empty (with probability $u(0)$),
- If the system contains one customer and that customer is of type two (with probability $u(1)/2$),
- If the system contains at least two customers and the two eldest customers are both of type two (with probability $[1 - u(0) - u(1)]\alpha/2$).

As a result, the probability that the first server is idle ($p_{I,1}$) equals

$$p_{I,1} = u(0) + \frac{1}{2}u(1) + \frac{1}{2}[1 - u(0) - u(1)]\alpha \ .$$

Due to the symmetry in the customer types, the probability that the second server is idle ($p_{I,2}$) equals the probability that the first server is idle:

$$p_{I,2} = p_{I,1} = u(0) + \frac{1}{2}u(1) + \frac{1}{2}[1 - u(0) - u(1)]\alpha \ . \tag{15}$$

This equation can be drastically simplified by applying the normalization condition of pgf's to (7), which yields

$$u(1) = \frac{2 - \alpha - \lambda - (2 - \alpha)u(0)}{1 - \alpha} \ . \tag{16}$$

Substituting (16) in (15) finally produces

$$p_{I,1} = p_{I,2} = 1 - \frac{\lambda}{2} \ ,$$

which thus highlights that class clustering has no impact on the idle probability of each server. This result can be explained as follows: the average number of customer arrivals of type i ($i = 1, 2$) in a random slot is equal to $\lambda/2$ (because both customer classes are equiprobable). In addition, this is, due to the steady state, also equal to the average number of customers that are served by server i in a random slot, which, in turn, is equal to the probability that server i is busy (due to the single-slot service times), which is the complementary probability of server i being idle.

5 Customer delay

We now turn to the analysis of the probability distribution of the delay customers incur in the system. More specifically, let C denote an arbitrary customer entering the system in steady state, and let S denote the slot during which C arrives. In the sequel, customer C will be referred to as the "tagged customer". We define the (discrete) delay d of customer C as the

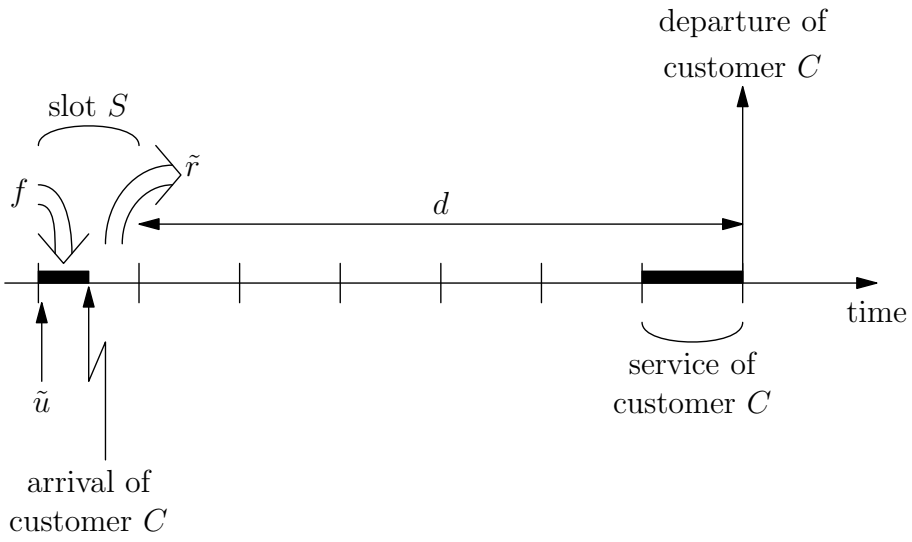

Figure 2: Time axis to illustrate the delay

total number of (full) slots between the arrival instant of C in the system and the departure time of C from the system, i.e., d indicates the number of slots between the end of slot S and the end of the slot during which C is actually being served (see Fig. 2).

Owing to the global FCFS service discipline used in the system, the delay d of the tagged customer C is equal to the number of slots required to serve all customers still in the system just after slot S, but to be served no later than C. In the next subsections, we first compute the pgf of this number of customers. Next, from this, we derive the pgf of d.

5.1 Customers to be served before the tagged customer

Let $\tilde{u}$ denote the system contents at the beginning of slot S, $\tilde{r}$ the number of "available servers" during slot S (equal to 1 or 2, with pgf $R(z)$ as defined in (2)), and f the number of customers entering the system during slot S but to be served before C (see Fig. 2). Then, the total number of customers to be served before the tagged customer C, still present in the system just after slot S, i.e., at the moment when the delay d of customer C starts running, is given by

$$v = (\tilde{u} - \tilde{r})^+ + f \ . \tag{17}$$

It is well-known from many previous papers e.g. [2, 13, 7, 14] that the pgf of the random variable f is given by

$$F(z) \triangleq E\left[z^f\right] = \frac{E(z) - 1}{(z-1)\,E'(1)} \ . \tag{18}$$

On the other hand, the independent nature of the arrival process (from slot to slot) implies that the probability distribution of $\tilde{u}$, i.e., the system occupancy at the beginning of the *arrival slot* of the tagged customer C, is identical to the probability distribution of the system occupancy at the beginning of an *arbitrary slot* in the steady state. This implies that the pgf of $\tilde{u}$

is equal to the function $U(z)$ determined earlier (see equation (7)). For the same reason, the random variables f and $\tilde{u}$ are also mutually independent. Putting all these elements together, we conclude that the pgf of v can be obtained as

$$V(z) \triangleq E[z^v] = E\Big[z^{(\tilde{u}-\tilde{r})^+}\Big] \cdot E\big[z^f\big] = \frac{U(z)}{E(z)} \cdot F(z) \ , \tag{19}$$

where, in the last step, we have used the steady-state version of equation (3), i.e., equation (3) for $k \to \infty$.

Using equations (7) and (18), we can derive from (19) the following explicit expression for $V(z)$:

$$V(z) = \frac{[E(z)-1]\,[u(0)(z+1-\alpha)+u(1)(1-\alpha)z]}{\lambda[z^2-(1-\alpha+\alpha z)E(z)]} \ . \tag{20}$$

5.2 Analysis of the delay

The delay d of customer C is nothing else than the number of slots required to remove the v customers in front of C just after slot S, together with customer C himself, from the system, i.e., the time needed to serve $v+1$ customers. If we denote by $\tilde{r}_j$ the number of "available servers" in the j-th slot following slot S, and by s_i the total number of customers that can be served during i consecutive slots (just after slot S), then it is not difficult to see that

$$s_i = \sum_{j=1}^{i} \tilde{r}_j \ , \tag{21}$$

with corresponding pgf

$$S_i(z) = R(z)^i \ , \tag{22}$$

with $R(z)$ as defined in equation (2).

The distribution of the delay d can then be obtained as follows. First, we express the tail distribution as

$$\text{Prob}[d > i] = \text{Prob}[s_i \leq v] = \sum_{n=0}^{\infty} \text{Prob}[s_i = n]\,\text{Prob}[v \geq n] \quad , \quad i \geq 0 \ , \tag{23}$$

with $\text{Prob}[s_i = n] = 0$, $n \leq i$. The reasoning behind this equation is that more than i slots are required to remove $v+1$ customers from the system, if and only if at most v customers can be served during i slots. We note that a similar approach was taken in [13] at the start of the analysis of the delay in a queueing system with variable service capacity, but here we present a somewhat more elegant method to arrive at closed-form results, based on the use of complex contour integration.

Specifically, in the above equation, we now represent $\text{Prob}[s_i = n]$ as a contour integral [6, 8]:

$$\text{Prob}[s_i = n] = \frac{1}{2\pi\imath} \oint_{C_x} S_i(x) x^{-n-1} dx = \frac{1}{2\pi\imath} \oint_{C_x} R(x)^i x^{-n-1} dx \ , \tag{24}$$

where $\imath$ denotes the imaginary unit and C_x a closed contour around the origin in the complex x-plane. At present time, this contour can be anywhere in the complex plane since $S_i(x)$ is polynomial and hence analytic in the whole complex plane.

On the other hand, we write $\text{Prob}[v \geq n]$ as

$$\text{Prob}[v \geq n] = \sum_{k=n}^{\infty} v(k) \ ,$$

where

$$v(k) \triangleq \text{Prob}[v = k] \ .$$

Equation (23) can then be rewritten as

$$\text{Prob}[d > i] = \frac{1}{2\pi\imath} \oint_{C_x} R(x)^i \sum_{n=0}^{\infty} \sum_{k=n}^{\infty} v(k) x^{-n-1} \ dx = \frac{1}{2\pi\imath} \oint_{C_x} R(x)^i \sum_{k=0}^{\infty} v(k) \sum_{n=0}^{k} x^{-n-1} \ dx \ . \tag{25}$$

Here, the sum over n is simply given by

$$\sum_{n=0}^{k} x^{-n-1} = \frac{x^{-k-1} - 1}{1 - x} \ ,$$

and the expression for $\text{Prob}[d > i]$ reduces to

$$\text{Prob}[d > i] = \frac{1}{2\pi\imath} \oint_{C_x} \frac{x^{-1} V(x^{-1}) - 1}{1 - x} \ R(x)^i \ dx \ .$$

Note that we have interchanged the order of the contour integral and the summations to arrive at (25). This is only allowed if the summations converge for all x on the contour (see e.g. [24]), i.e., in our case, we have to assume that

$$|x^{-1}| < R_V,$$

for all $x \in C_x$, with R_V the radius of convergence of V. After a change of integration variable from x to $y = x^{-1}$ and adaptation of the contour C_x into its image C_y, but still integrating along C_y in counter-clockwise sense (which yields an extra factor -1), this can be rewritten as

$$\text{Prob}[d > i] = \frac{1}{2\pi\imath} \oint_{C_y} \frac{yV(y) - 1}{y(y-1)} \ [R(y^{-1})]^i \ dy \ . \tag{26}$$

Here, $|y| < R_V$ for all $y \in C_y$.

Now, let $D(z)$ denote the pgf of the delay d, then it is easily seen that

$$\sum_{i=0}^{\infty} z^i \ \text{Prob}[d > i] = \frac{D(z) - 1}{z - 1} \ .$$

Multiplying both sides of equation (26) with z^i and summing over all nonnegative values of i, we therefore obtain the following result for $D(z)$:

$$D(z) = 1 + (z-1)\frac{1}{2\pi\imath}\oint_{C_y} \frac{yV(y)-1}{y(y-1)(1-zR(y^{-1}))}\ dy\ . \tag{27}$$

In order for this summation over i to converge for all z inside the open unit disk, we require that $1 < |y|$ for all $y \in C_y$. This leads to the final condition for the location of the contour C_y:

$$1 < |y| < R_V\ ,$$

for all $y \in C_y$.

Formula (27) expresses the pgf of the delay of an arbitrary customer in terms of known quantities only, albeit in a not very transparent way. One way of evaluating the complex contour integral in equation (27) is applying Cauchy's residue theorem from complex analysis [8], which states in general that a contour integral of the form

$$\frac{1}{2\pi\imath}\oint_{C_y} g(y)\ dy$$

can be expressed as the sum of the residues of the integrand $g(y)$ in the singularities of $g(y)$ which are located inside the closed contour C_y. In this particular case, the integrand is given by

$$g(y) = \frac{yV(y)-1}{y(y-1)(1-zR(y^{-1}))} = \frac{y(yV(y)-1)}{(y-1)(y^2-\alpha zy-(1-\alpha)z)}$$

and the singularities of $g(y)$ (possibly) inside the closed contour C_y are $\{y \mid 1-zR(y^{-1})=0\}$ (note that $y=1$ is a removable singularity since $V(1)=1$).

The solutions of the (quadratic) equation $1-zR(y^{-1})=0$ are given by

$$y_1 = \frac{\alpha z + \sqrt{\alpha^2 z^2 + 4(1-\alpha)z}}{2} \text{ and } y_2 = \frac{\alpha z - \sqrt{\alpha^2 z^2 + 4(1-\alpha)z}}{2}\ , \tag{28}$$

where y_1 and y_2 are, in fact, shorthand notations for $y_1(z)$ and $y_2(z)$. It can be proved that $|y_i(z)| < 1$ for $|z| < 1$, hence these singularities lie inside the closed contour C_y for all z inside the open unit disk. The residues of $g(y)$ in $y=y_1$ and in $y=y_2$ are given by

$$Res_{y_1}[g(y)] = \lim_{y\to y_1}(y-y_1)\ g(y) = y_1\ \frac{y_1V(y_1)-1}{(y_1-1)(y_1-y_2)} \tag{29}$$

and

$$Res_{y_2}[g(y)] = \lim_{y\to y_2}(y-y_2)\ g(y) = y_2\ \frac{y_2V(y_2)-1}{(y_2-1)(y_2-y_1)}\ . \tag{30}$$

Putting all elements together, we can derive from equation (27) the following expression for $D(z)$:

$$D(z) = \frac{z-1}{y_1-y_2}\left[\frac{y_1^2\ V(y_1)}{y_1-1} - \frac{y_2^2\ V(y_2)}{y_2-1}\right]\ , \tag{31}$$

which can also be expressed as

$$D(z) = \frac{z}{\sqrt{\alpha^2 z^2 + 4(1-\alpha)z}} \{(1-\alpha+\alpha z)[V(y_1) - V(y_2)] + y_1 V(y_2) - y_2 V(y_1)\} \quad . \tag{32}$$

Equations (31) and (32) represent closed-form expressions for the pgf $D(z)$ as soon as y_1 and y_2 are replaced by the expressions in (28) and the known form of the function $V(z)$ is taken from (20).

5.3 Special case: geometric arrivals

Let us consider again the special case whereby the number of arrivals per slot has a geometric distribution with mean value λ. As explained above (see equation (9)), the pgf $U(z)$ of the system occupancy at the beginning of an arbitrary slot then reduces to

$$U(z) = \frac{1-z_0}{z-z_0} \ ,$$

where z_0 is given in equation (10). From equations (18) and (19) it easily follows that the pgf $V(z)$ is then also equal to

$$V(z) = \frac{1-z_0}{z-z_0} \ , \tag{33}$$

which means that the random variable v is geometrically distributed with parameter $1/z_0$, i.e.,

$$v(k) = \left(1 - \frac{1}{z_0}\right)\left(\frac{1}{z_0}\right)^k \quad . \tag{34}$$

This circumstance simplifies the direct computation of the delay distribution considerably: starting from equation (23) we immediately get

$$\text{Prob}[d > i] = \sum_{n=0}^{\infty} \text{Prob}[s_i = n]\,[1/z_0]^n \ ,$$

for all $i \geq 0$. This can be further expressed as

$$\text{Prob}[d > i] = S_i(1/z_0) = [R(1/z_0)]^i = \left(\frac{\alpha}{z_0} + \frac{1-\alpha}{z_0^2}\right)^i \ , \tag{35}$$

according to equations (22) and (2). The above result shows that, in case of geometric arrivals, just as the system occupancy, the delay has a geometric distribution as well. More specifically, whereas the system occupancy is geometrically distributed with parameter $1/z_0$, the delay is geometrically distributed with parameter $R(1/z_0)$, i.e., with pmf

$$d(i) \triangleq \text{Prob}[d = i] = [1 - R(1/z_0)]\ [R(1/z_0)]^{i-1} \ , \tag{36}$$

for all $i \geq 1$, and pgf

$$D(z) = \frac{z[1 - R(1/z_0)]}{1 - zR(1/z_0)} \ . \tag{37}$$

The mean customer delay $E[d]$ is given by

$$E[d] = \frac{1}{1 - R(1/z_0)} \ . \tag{38}$$

From the definition of z_0 as a zero of the denominator of (7), it is not difficult to show that the parameter $R(1/z_0)$ can also be expressed as

$$R(1/z_0) = \frac{1}{E(z_0)} = 1 + \lambda - \lambda z_0 \ . \tag{39}$$

It follows that the mean delay can be written as

$$E[d] = \frac{1}{\lambda(z_0 - 1)} \ , \tag{40}$$

in full agreement with equations (13) and (14), i.e., with Little's theorem.

Finally, we note that the pgf $D(z)$ of the delay in case of geometric arrivals, as given in equation (37), can also be obtained by introducing the specific form of the pgf $V(z)$, as given by (33), in either of the equations (31) or (32) and using the definitions of y_1 and y_2 in (28). The calculations needed to show this are somewhat tedious, but straightforward, and are therefore omitted here. We remind the reader that in case the random variable v is not geometrically distributed (which would in general be the case for non-geometric arrivals) the simplified delay analysis presented in this subsection is not applicable, and the results (31) and (32) are the only ones available.

6 Discussion of results and numerical examples

In this section, we discuss the results obtained in the previous sections, both from a qualitative perspective and by means of some numerical examples.

The first interesting result obtained is the form of the stability condition (6),

$$\lambda < 2 - \alpha \ ,$$

which shows that the maximum achievable throughput of this system, expressed in customers per slot, is very directly determined by the degree of class clustering in the arrival process as described by the cluster parameter α. For this specific model, the formula is remarkably simple and shows that the achievable throughput decreases linearly with the cluster parameter α. As α can take values between 0 and 1, the maximum throughput can vary between (nearly) 2 customers per slot and (nearly) 1 customer per slot. It is interesting to look at the extreme values $\alpha = 0$ and $\alpha = 1$. If the cluster parameter is equal to zero, then the types of two consecutive customers are always opposite, and one type of customers can never block the

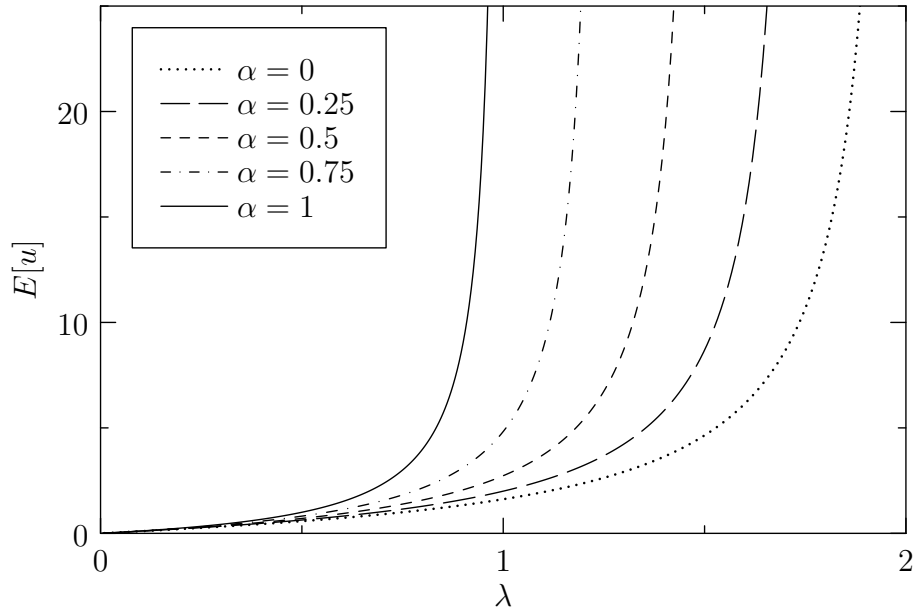

Figure 3: Mean system contents versus the mean arrival rate for various values of the cluster parameter

other type; in this case both servers A and B are active as soon as at least two customers are present in the system, i.e., the system is work-conserving and behaves as a regular queue with two identical servers able to serve all customers. However, as soon as some amount of "class clustering" appears in the arrival stream, the achievable throughput is affected, according to equation (6). In the extreme case where the cluster parameter is equal to 1, all customers belong to the same class and only one of the two servers is actually being used by the arrival stream; in this case, the system behaves as a single-server queue and the throughput can never exceed 1 customer per slot.

These results show that the presence of "class clustering" in the arrival stream of a multiclass queue with dedicated servers and "global FCFS" service discipline can actually be devastating for the performance of the queue, and we believe that this phenomenon has been largely overlooked in the regular queueing literature. Another way of looking at this phenomenon is to rewrite the inequality (6) as

$$\lambda + \alpha < 2 \ , \tag{41}$$

which seems to say that the actual traffic intensity (λ) and the cluster parameter (α) are equally important with respect to the stability of the queue: you can afford more load only if you can decrease the class clustering of the arrival stream, i.e., the class clustering appears to represent some kind of additional or virtual load to the system. In this sense, the quantity $\lambda + \alpha$ could be considered as some kind of equivalent traffic intensity of the system.

For the case of geometric arrivals, as discussed in subsections 3.3 and 5.3, we show some numerical results in figures $3 - 7$.

Fig. 3 shows the mean system contents $E[u]$ versus the mean arrival rate λ, for various

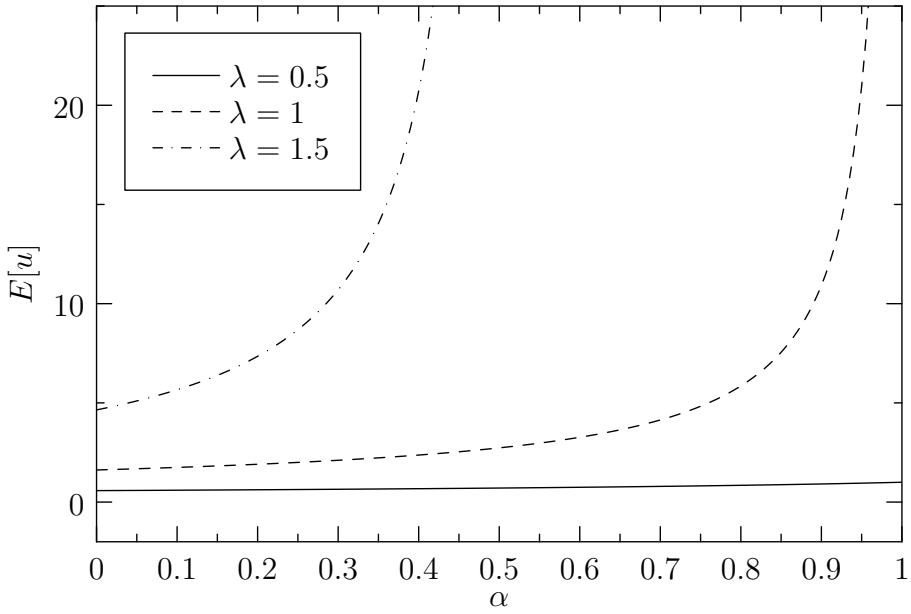

Figure 4: Mean system contents versus the cluster parameter for various values of the mean arrival rate

values of the cluster parameter α. The figure clearly illustrates the great and direct (negative) impact of "class clustering" on the average number of customers in the system, for any given arrival intensity lower than 1. More generally, it also shows the shrinking stability region of the system, as the degree of class clustering increases. We note that the value $\alpha = 0.5$ represents the case where the types of consecutive customers in the arrival stream are independent. Our results prove that neglecting the correlation between the types of consecutive customers may lead to either serious underestimation or overestimation of the mean system occupancy.

In Fig. 4, we have plotted the mean system contents $E[u]$ versus the cluster parameter α, for given values of the arrival rate λ. The figure shows that for lightly loaded systems (e.g. $\lambda = 0.5$ in the figure) the influence of class clustering is negligible. This is also intuitively clear: the demand of the arrival stream, in such a case, is considerably less than the traffic that can be handled by 1 server, and therefore, the question of whether the second server is also active or not — which is determined by the amount of class clustering — is not very relevant. However, as soon as the arrival rate λ exceeds the value 1, the cluster parameter α becomes important. Specifically, the average queue size can even grow without bound when α reaches the value $2 - \lambda$.

Fig. 5 shows the tail probability $\text{Prob}[u > i]$, which can be considered as an approximate value for the loss probability in a system with finite storage capacity equal to i places, versus the value of i, for a given value $\lambda = 1$ and various values of the cluster parameter α. The results in this figure can be used, for instance, for dimensioning purposes of the required buffer size to achieve a prescribed loss ratio. As an example, let us assume a target loss ratio of 10^{-4}, then the graphs in Fig. 5 show that the required buffer size depends very strongly on the

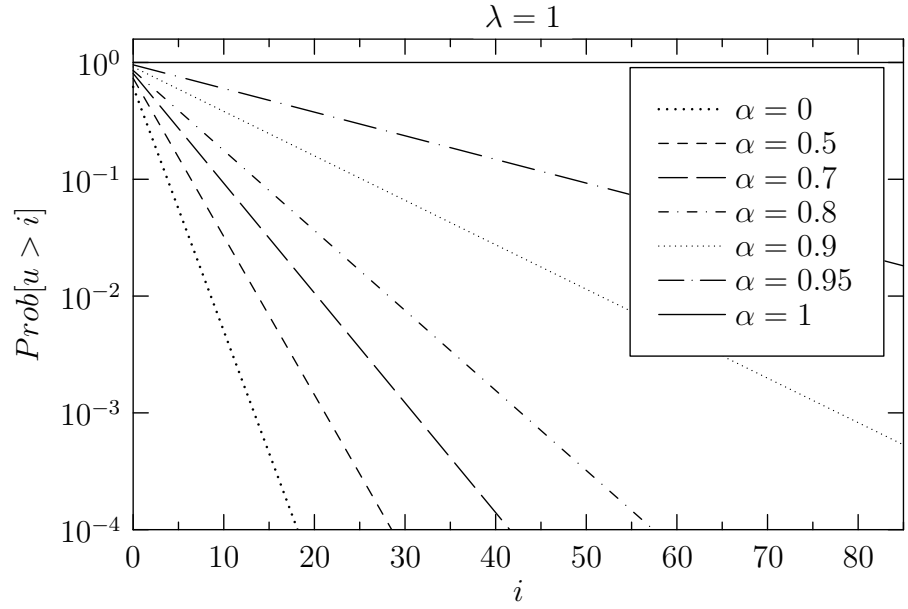

Figure 5: Tail probability of the system contents for a given arrival rate of 1 and various values of the cluster parameter

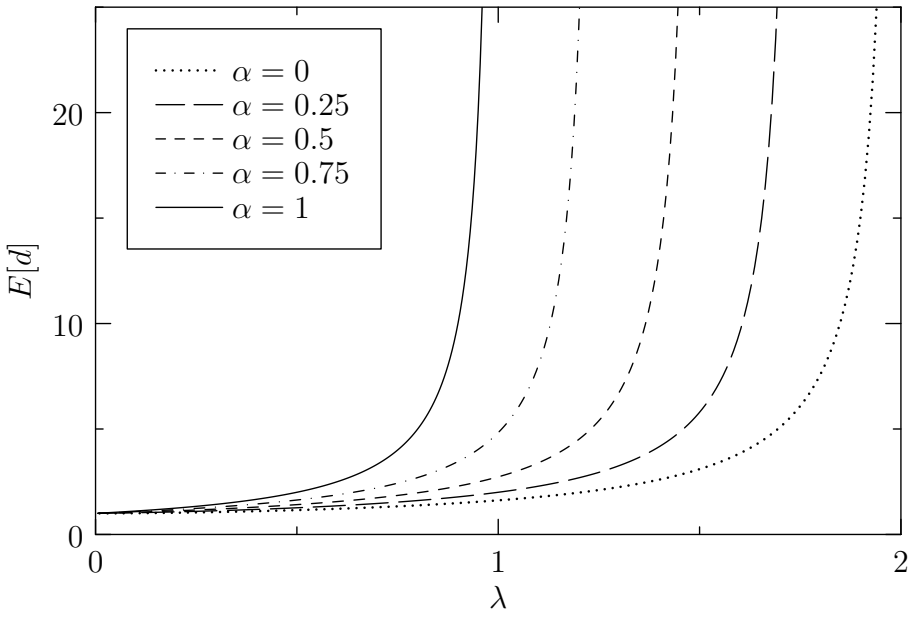

Figure 6: Mean delay versus the mean arrival rate for various values of the cluster parameter

cluster parameter α: for $\alpha = 0$, a storage capacity of 18 is sufficient; $\alpha = 0.5, 0.7, 0.8, 0.9$ and 0.95 require a buffer size of 29, 42, 58, 105, 197 respectively, whereas for $\alpha = 1$ the system is unstable and a loss ratio of 10^{-4} is not even achievable.

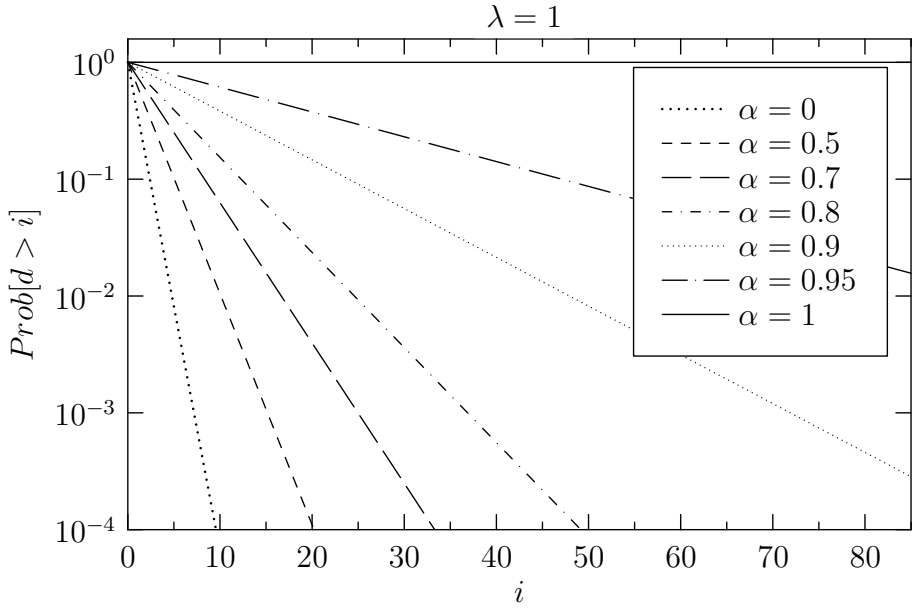

Figure 7: Tail probability of the delay for a given arrival rate of 1 and various values of the cluster parameter

Fig. 6 shows the mean delay $E[d]$ versus the mean arrival rate λ, for various values of the cluster parameter α. Again the detrimental impact of class clustering on the performance of the system is clearly reflected in the figure.

Next, in Fig. 7 we display the tail probability of the delay, for a given value $\lambda = 1$ and various values of the cluster parameter α. The graph illustrates that the quantiles of the delay distribution are very dependent on the cluster parameter α. For instance, the 10^{-4}-quantiles are given by 10, 21, 34, 50, 96 and 189 respectively for $\alpha = 0, 0.5, 0.7, 0.8, 0.9$ and 0.95. Comparison of Figs. 5 and 7 also clearly reveals the relationship between the decay rates of the system content distribution and the delay distribution, which are given by $1/z_0$ (see equation (12)) and $R(1/z_0) = \alpha/z_0 + (1-\alpha)/z_0^2$ (see equation (35)) respectively. For $\alpha = 0$, the decay rate of the delay is given by $(1/z_0)^2$, which means that, in a semi-logarithmic graph, the slope of the tail distribution of the delay is twice the slope of the tail distribution of the system occupancy, reflecting the fact that, in this case, the system behaves as a regular two-server queue and customers can be removed from the system at the rate of two per slot. For $\alpha \to 1$, on the other hand, the decay rate of the delay is simply given by $1/z_0$, i.e., equal to the decay rate of the system occupancy, which stems from the fact that the system reduces to a single-server queue in this case, whereby one single customer can be removed per slot. For intermediate values of α, the slope of the curves for the tail of the delay distribution is between 1 and 2 times the slope of the system occupancy curves.

Finally, we compare our system with global FCFS with a related system with "partial FCFS", i.e., with two separate queues for each type of service. In the latter system, customers of distinct types cannot block each other anymore. It can still happen that only one of the

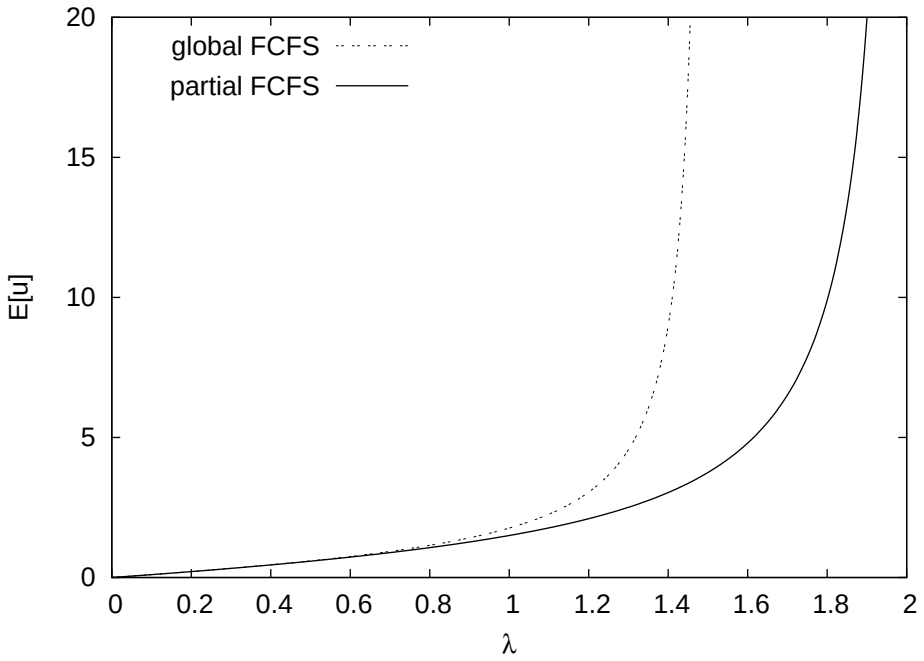

Figure 8: Mean system contents versus the mean arrival rate for global FCFS as well as for partial FCFS

servers is active even if the entire system contains more than one customer (all of the same type, in such a case). It is important to realize that a system with partial FCFS is very hard to analyze, because of the correlation that exists (i) between the types of subsequent customer arrivals, even across slot boundaries, which will lead to a correlated (marginal) arrival process in the two "partial FCFS" queues and (ii) between the number of customers that arrive in the same slot in these two queues. Especially this latter effect is extremely hard to deal with, since in such a scenario we have two non-independent queues. However, the case of $\alpha = 0.5$ and a Poisson distribution for the number of total customer arrivals in a random slot (i.e. $E(z) = e^{\lambda(z-1)}$) is a welcome exception. Indeed, in this case, for $\alpha = 0.5$, the type of consecutive customers is an uncorrelated process (therefore no correlation across slot boundaries exists). In addition, an aggregate Poisson arrival process that is decomposed in a probabilistic way, leads to two independent Poisson processes. As a result, the number of arrivals in queue 1 in a random slot is independent of and identically distributed as the number of arrivals in queue 2 in that slot, and it has a Poisson distribution with parameter $\lambda/2$. As a consequence, the mean system contents in the individual queues (denoted by $\mathrm{E}[\tilde{u}_1]$ and $\mathrm{E}[\tilde{u}_2]$) have the same value, and the total mean system content ($\mathrm{E}[\tilde{u}]$) is then equal to $2\mathrm{E}[\tilde{u}_1]$. The mean system content $\mathrm{E}[\tilde{u}_1]$ is given by the well-known formula for a discrete-time single-server system with service time of one slot (see e.g. [3]), which finally yields the following formula for $\mathrm{E}[\tilde{u}]$:

$$\mathrm{E}[\tilde{u}] = \frac{\lambda(4-\lambda)}{2(2-\lambda)} \ .$$

In Fig. 8, the mean system content is depicted versus the mean arrival rate λ, both for the global FCFS and for the partial FCFS service policy. As anticipated, we observe that the blocking effect stemming from global FCFS leads to an undeniable decrease in performance, reflected by a shrinking stability region and a larger mean system content (and customer delay).

7 Conclusions

In this paper, we have studied a dual-class, two-server queue with class-dedicated servers in discrete time, operating under the global FCFS service discipline, assuming independent arrivals from slot to slot with a simple first-order Markovian class clustering model. The system is relatively simple so as to allow for an analytical solution, but yet contains all the important elements needed for a conceptual study of the effect of "global FCFS" on this type of queue. We emphasize that we have succeeded in deriving explicit closed-form formulas for the idle probability of each server and explicit semi-analytic formulas for the pgf's of the system occupancy and the delay, under general assumptions with respect to the arrival statistics. For the special case of geometric arrivals, we have even been able to obtain explicit closed-form expressions for the pmf's, the mean values and the tail distributions of system occupancy and delay. The results reveal the very direct and great influence of the degree of "class clustering" in the arrival stream on the stability and the main performance measures of the system. Only the idle probabilities of each server remain unaltered. We believe that this is the main qualitative conclusion of the study.

In general, only few studies have focused on the phenomenon of class clustering in the context of multiclass queueing systems, and this paper shows that the effect of class clustering may be very important, possibly not only in queues with class-dependent servers and global FCFS, but also in other queueing situations whereby the service mechanism is sensitive to the order of service of customers of different classes. For instance, we expect that class clustering may have substantial effects on the performance of priority queues: low priority customers might suffer excessive delays when a heavy clustering of high priority customers exists. Also, class clustering might be significant in queueing models where the lengths of the service times depend on the way customers of different types succeed each other. Consider for instance a machine with two modes, corresponding to two product types that can be processed. When the next customer is of another type, the machine has to make a switch, which takes some time and thus leads to a longer total processing time.

The model examined in this paper can be generalized in various directions. To start with, the symmetric two-state Markov chain to model the types of subsequent customers can be relaxed by considering an asymmetric two-state Markov chain. Such a model will be harder to solve, as the system description will have to keep track of the type of the eldest customer in the queue. The number of customers that can be served depends on the equalness of the types of the two eldest customers. Precisely due to the symmetry in the two-state Markov chain in this paper, the type equalness does not depend on the type of the eldest customer, so it is not necessary to keep track of the type of the eldest customer. Keeping track of the type of the eldest customer in case of an asymmetric two-state Markov chain will therefore require a more

complicated state description, leading to a harder but still tractable analysis. Next, more than two customer classes can be studied in the future. In case of more than two customer classes, the Markov chain that assigns a specific type to subsequent customers will contain more than two states. Therefore, the definition of our cluster parameter α will not be valid anymore, since a single parameter is no longer sufficient to capture the entire customer type assignment process. Note that the state space of the model grows exponentially with the number of classes. Due to this extended state space, the model will be harder to analyze, but it might still be possible for a relatively low number of classes. In addition, more general service-time distributions can be considered than the simple deterministic one-slot-per-customer model studied in this paper. For service times distributions that possess the memoryless property (i.e., a geometric distribution for the customer service times, or some "phase-type" extension), the analysis will probably be feasible. Nonetheless, even for the relatively simple geometric scenario it will be considerably more difficult, due to a new phenomenon that may now occur: customers (of a different type) can now "overtake" each other. Indeed, if both servers contain a customer, then the customer that has first entered its server is not necessarily the first one to leave. This mechanism leads to a more difficult and messy system description (and a larger state space), where one even may have to keep track of the type of a customer that has already left the system. In case of generally distributed service times, we do not think that the system can be explicitly solved. Note that even the basic discrete-time two-server queueing system with generally distributed service times has not been brought to a closed-form solution up to now. We plan to tackle several of these generalizations in future work.

We would like to conclude by emphasizing that we have opted to keep the modelling assumptions in this contribution as simple as possible, in order to highlight the impact of class clustering on customer blocking. We think that they have proven to be an adequate choice for such a purpose.

Acknowledgment The authors would like to thank the anonymous referees and the editor for their constructive suggestions, which led to a considerable improvement of this paper.

References

[1] P. Beekhuizen and J. Resing. Performance analysis of small non-uniform packet switches. *Performance Evaluation*, 66:640–659, 2009.

[2] H. Bruneel. Buffers with stochastic output interruptions. *Electronics Letters*, 19:735–737, 1983.

[3] H. Bruneel and B.G. Kim. *Discrete-time models for communication systems including ATM*. Kluwer Academic, Boston, USA, 1993.

[4] D. Claeys, J. Walraevens, K. Laevens, and H. Bruneel. Analysis of threshold-based batch-service queueing systems with batch arrivals and general service times. *Performance Evaluation*, 68(6):528–549, 2011.

[5] D. Fiems and H. Bruneel. A note on the discretization of Little's result. *Operations Research Letters*, 30:17–18, 2002.

[6] F. Flanigan. *Complex variables; harmonic and analytic functions*. Dover Publications, New York, USA, 1983.

[7] P. Gao, S. Wittevrongel, and H. Bruneel. Discrete-time multiserver queues with geometric service times. *Computers and Operations Research*, 31:81–99, 2004.

[8] M.O. Gonzáles. *Classical complex analysis*. Marcel Dekker, New York, USA, 1992.

[9] F.N. Gouweleeuw and H.C. Tijms. Computing loss probabilities in discrete-time queues. *Operations Research*, 46:149–154, 1998.

[10] A. Kesselman, K. Kogan, and M. Segal. Improved competitive performance bounds for CIOQ switches. *Algorithmica*, 63(1-2):411–424, 2012.

[11] H.S. Kim and N.B. Schroff. Loss probability calculations and asymptotic analysis for finite buffer multiplexers. *IEEE/ACM Trans. on Networking*, 9:755–768, 2001.

[12] L. Kleinrock. *Queueing systems, part I*. Wiley, New York, USA, 1975.

[13] K. Laevens and H. Bruneel. Delay analysis for discrete-time queueing systems with multiple randomly interrupted servers. *European Journal of Operations Research*, 85:161–177, 1995.

[14] T. Maertens, J. Walraevens, and H. Bruneel. A modified HOL priority scheduling discipline: performance analysis. *European Journal of Operations Research*, 180:1168–1185, 2007.

[15] A. Mandelbaum and M.I. Reiman. On pooling in queueing networks. *Management Science*, 44:971–981, 1998.

[16] D. Ngoduy. Derivation of continuum traffic model for weaving sections on freeways. *Transportmetrica*, 2:199–222, 2006.

[17] R. Nishi, H. Miki, A. Tomoeda, and K. Nishinari. Achievement of alternative configurations of vehicles on multiple lanes. *Physical Review E*, 79:066119, 2009.

[18] E. Oki, N. Kitsuwan, and R. Rojas-Cessa. Performance analysis of clos-network packet switch with virtual output queues. *IEICE Transactions on Communications*, E94B(12):3437–3446, 2011.

[19] D. Shah, J.N. Tsitsiklis, and Y. Zhong. Optimal scaling of average queue sizes in an input-queued switch: an open problem. *Queueing Systems*, 68(3-4):375–384, 2011.

[20] B. Steyaert and H. Bruneel. *Accurate approximation of the cell loss ratio in ATM buffers with multiple servers*, volume 1 of *Performance Modelling and Evaluation of ATM Networks*, pages 285–296. Chapman & Hall, London, 1995.

[21] N.M. Van Dijk and E. Van der Sluis. To pool or not to pool in call centers. *Production and Operations Management*, 17:1–10, 2008.

[22] T. Van Woensel and N. Vandaele. Empirical validation of a queueing approach to uninterrupted traffic flows. *4OR, A Quarterly Journal of Operations Research*, 4:59–72, 2006.

[23] T. Van Woensel and N. Vandaele. Modeling traffic flows with queueing models: A review. *Asia-Pacific Journal of Operational Research*, 24:435–461, 2007.

[24] B. Vinck. *System content and sojourn time in discrete-time queueing systems*. PhD thesis, Ghent University, 2006.

[25] N. Wu. Modelling blockage probability and capacity of shared lanes at signalized intersections. *Procedia - Social and Behavioral Sciences*, 16:481–491, 2011.

[26] H. Yu, S. Ruepp, and M.S. Berger. Enhanced first-in-first-out-based round-robin multicast scheduling algorithm for input-queued switches. *IET Communications*, 5(8):1163–1171, 2011.

[27] Y.H. Zhang, X.G. Dong, S.Q. Gan, and W.M. Zheng. Model of network-on-chip routers and performance analysis. *IEICE Electronics Express*, 8(13):986–993, 2011.